Inventors:

**Donald WILSON**
115 Havilands Lane
White Plains, NY 10605

**Anthony H. HANDAL**
3 Blue Chip Lane
Westport CT 06880

**Michael LESSAC**

# SPEECH RECOGNITION METHOD

## TECHNICAL FIELD

The present invention relates to speech recognition technology of the type typically embodied in speech recognition software typically implemented on personal computer systems.

## BACKGROUND

In 1964, a group of computer scientists marveled over the new computer being delivered into the computer center at the Courant Institute for Mathematical Sciences at New York University. The machine was the latest introduction from Control Data Corporation, a Model CDC 6600, whose speed and memory capacity far outstripped the 7.094 K random access memory capacity of the now humble IBM 7094 that it replaced. In a portent of things to come, they little suspected that IBM, within months, would obsolete the long heralded CDC 6600 with its IBM 360, a machine which, incredibly, had an unheard-of 360 K of RAM, all built with discrete components, a conservative move in the face of concerns about the reliability of the then-new integrated circuit technology. This impressive machine came to be housed in a room about eighteen feet square, and surrounded by ten or so air conditioners necessary to keep the system from overheating and failing. A half dozen tape decks, nearly a meter across and as

tall as a man, and several key punch machines, the size of garment outer table serving machines completed the installation.

Thirty-five years later, changes in technology have been remarkable. Tiny laptop computing devices fly at speeds a thousand times that of those early powerhouse computers and boast thousands of times the memory. Instead of huge reels of recording tape, hard disks with capacities on the order of eighteen GB are found in those same laptop computing devices. These devices, with their huge memory and computing capabilities, move as freely in the business world as people, under the arm, in a bag, or on the lap of a businessman flying across the ocean. No doubt, this technology lies at the foundations of the most remarkable, reliable and completely unanticipated bull market in the history of business.

Just as certainly, the future holds the promise of similar progress.

Notwithstanding the gargantuan magnitude of the progress made in computing during the last third of the 20th century, the world of computing has been largely self-contained. The vast majority of all computing tasks involved

computers talking to other computers, or otherwise communicating through use of electrical input signals whose characteristics are substantially absolutely determined. In this respect, computers are completely unlike the humans which they serve, humans whose communications vary in almost infinite ways, regardless of the method of communication, be it voice, writing, or other means. If computing is to continue to make progress, computers must become integrated into usual human communications modalities.

And, indeed, this is already happening. From a slow start at becoming an important factor in the marketplace about a decade ago, speech recognition technology holds just such a promise. A human voice interface to a computer represents what should be probably the most ideal evolutionarily defined modality for the human-computer communications interface. While humans customarily write, gesture and, to a limited extent, use other communications modes, voice communication remains predominant. This is not surprising, insofar as speech has evolved in human beings probably for many millions of years. This is believed to be the case because even relatively primitive forms of life have fairly highly developed "speech" characteristics. For example, much work has been done in the study of the use of sounds to communicate various

items of information by whales. Likewise, scientists have identified and cataloged uniform global communications patterns in chimpanzees.

In view of the highly natural nature of communication by speech, a direct result of its having evolved over such a large fraction of the history of the species, speech communications impose an extremely low level of cognitive overhead on the brain, thus providing a facile communications interface while allowing the brain to perform a number of other functions simultaneously. We see this in everyday life. For example, people engaged in sports activities routinely combine complex physical tasks, situational analysis, and exchange of information through speech, sometimes simultaneously transmitting and receiving audible information, while doing all of these other tasks.

Clearly, the mind is well adapted to simultaneously control other tasks while communicating and receiving audible information in the form of speech. It is thus no surprise that virtually every culture on earth has devised its own highly sophisticated audible language.

In view of the above, it is thus, easily understood why voice recognition

technology has come to be the Holy Grail of computing. While useful work began to be done with this technology about ten years ago, users only obtained performance which left much to hope for. Individual quirks, regional pronunciations, speech defects and impediments, bad habits and the like pepper virtually everybody's speech to some extent. And this is no small matter. Good speech recognition requires not only good technology, it also requires recognizable speech.

Toward this end, speech recognition programs generally have an error correction dialog window which is used to train the system to the features of an individual user's voice, as will be more fully described below. The motivation behind such technique is apparent when one considers and analyzes the schemes typically used in speech recognition systems.

Early on, speech recognition was proposed through the use of a series of bandpass filters. These proposals grew out of the use of spectrographic analysis for the purpose of speaker identification. More particularly, it was discovered that if one made a spectral print of a person saying a particular word, wherein the x-axis represented time and y-axis represented frequency, with the intensity

of sound at the various frequencies being displayed in shades of gray or in black and white, the pattern made by almost every speaker was unique, largely as a function of physiology, and speakers could be identified by their spectrographic "prints". Interestingly enough, however, the very diversity which this technique showed suggested to persons working in the field the likelihood that commonalities, as opposed to differences, could be used to identify words regardless of speaker. Hence the proposal for a series of bandpass filters to generate spectrographs for the purpose of speech recognition.

While such an approach was logical given the state of technology in the 1960s, the problems were also apparent. Obtaining high-quality factors or "Q" in electrical filters comprising inductors and capacitors is extremely difficult at audio frequencies.

This is due to a number of factors. First of all, obtaining resonance at these frequencies necessitates the use of large capacitors and inductors. Such components, in the case of capacitors have substantial resistance leak through. In the case of inductances, large values of inductance are required, thus requiring large lengths of wire for the windings and, accordingly, high

resistance. The result is that the selectivity of the filters is extremely poor and the ability to separate different bandpasses is compromised. Finally, the approach was almost fatally flawed, from a mass-market standpoint, by the fact that these tuned electrical circuits were very large and mechanically cumbersome, as well as very expensive.

However, in the late 1960's, electrical engineers began to model the action of electrical circuits in the digital domain. This work was done by determining, using classical analytic techniques, the mathematical characteristics of the electrical circuit, and then solving these equations for various electrical inputs. In the 1970's, it was well understood that the emerging digital technology was going to be powerful enough to perform a wide variety of computing tasks previously defaulted to the analog world. Thus, it was inevitable that the original approaches to voice recognition through the concept of using banks of tuned circuits would eventually come to be executed in the digital domain.

In a typical speech recognition system, an acoustic signal received by a microphone is input into a voice board which digitizes the signal. The computer then generates a spectrogram which, for a series of discrete time intervals,

records those frequency ranges at which sound exists and the intensity of sound in each of those frequency ranges. The spectrogram, referred to in the art as a token, is thus a series of spectrographic displays, one for each of a plurality of time intervals which together form an audible sound to be recognized. Each spectrographic display shows the distribution of energy as a function of frequency during the time interval. In a typical system, sampling rates of 6,000 to 16,000 samples per second are typical, and are used to generate about fifty spectrum intervals per second for an audible sound to be recognized.

In a typical system, quantitative spectral analysis is done for seven frequency ranges, resulting in eight spectral parameters for each fiftieth of a second, or spectral sample period. While the idea that a spectral analysis over time can be a reliable recognition strategy may be counterintuitive given the human perspective of listening to envelope, tonal variation and inflection, an objective view of the strategy shows that exactly this information is laid out in an easy to process spectral analysis matrix.

Based on the theoretical underpinnings of the above recognition strategy, development of a speech recognition system involves the input of vocabulary

into the hard drive of a computer in the form of the above described spectral analysis matrix, with one or more spectral analysis matrices for each word in the vocabulary of the system. These matrices then serve as word models.

In more advanced systems (such as those using so-called "natural" speech, that is continuous strings of words, the natural tendency of speakers to, on occasion, blend the end of one word into the beginning of another, and less frequently to separate words into two parts, sometimes with association of the parts with different words) models are also developed for these artifacts of the language to be recognized.

Once broken down into a spectral picture over time of frequency energy distributions, recognition of speech is reduced to comparison of known spectral pictures for particular sounds to the sound to be recognized, and achieving recognition through the determination of that model which best matches the unknown speech sound to be recognized. But this picture, while in principle correct, is an unrealistic simplification of the problem of speech recognition.

After a database of word models has been input into the system, comparison of

an audible sound to the models in the database can be used as a reliable means for speech recognition. However, there are many differences in the speech patterns of users. For example, different speakers speak at different rates. Thus, for one speaker, a word they take a certain period of time, while for another speaker, the word they take a longer period of time. Moreover, different speakers have voices of different pitch. In addition, speakers may give different inflections, emphasis, duration and so forth to different syllables of a word in different ways, depending on the speaker. Even a single speaker will speak in different ways on different occasions.

Accordingly, effective speech recognition requires normalization of spoken sounds to word and phrase models in the database. In other words, the encoded received sound or token must be normalized to have a duration equal to that of the model. This technology is referred to as time aligning, and results in stretching out or compressing the spoken sound or word to fit it against the model of the word or sound with the objective of achieving the best match between the model and the sound input into the system. Of course, it is possible to leave the sound unchanged and stretch or compress the model.

In accordance with existing technology, each of the spectral sample periods for the sound to be recognized are compared against the corresponding spectral sample periods of the model which is being rated. The cumulative score for all of the sample periods in the sound against the model is a quality rating for the match. In accordance with existing technology, the quality ratings for all the proposed matches are compared and the proposed match having the highest quality rating is output to the system, usually in the form of a computer display of the word or phrase.

However, even this relatively complex system fails to achieve adequate quality in the recognition of human speech. Accordingly, most commercial systems, do a contextual analysis and also require or strongly recommend a period of additional training, during which the above matching functions are performed with respect to a preselected text. During this process, the model is appended to take into account the individual characteristics of the person training the system. Finally, during use, an error correction dialog box is used when the user detects an error, inputs this information into the system and thus causes the word model to become adapted to the user's speech. This supplemental training of the system may also be enhanced by inviting the user, during the error correction

dialog to speak the word, as well as other words that may be confused with the word by the system, into the system to further train the recognition engine.

As is apparent from the above discussion, the development of speech recognition systems has centered on assembling a database of sound models likely to have a high degree of correlation to the speech to be recognized by the speech recognition engine. Such assembly of the database takes two forms. The first is the input of global information using one or more speakers to develop a global database. The second method in which the database is assembled is the training of the database to a particular user's speech, typically done both during a training session with preselected text, and on an *ad hoc* basis through use of the error correction dialog window in the speech recognition program.

## SUMMARY OF THE INVENTION

In accordance with the invention, the performance of the speech recognition software is improved by focusing in on the user, as opposed to the software. In particular, the invention has its objective improvement of the speech patterns of persons using the software. The result is enhanced performance, with the bonus of voice training for the user. Such training is of great importance. For example,

salesman, lawyers, store clerks, mothers dealing with children and many others rely heavily on oral communications skills to accomplish daily objectives. Nevertheless, many individuals possess poor speaking characteristics and take this handicap with them to the workplace and throughout daily life.

Perhaps even more seriously, speech defects, regionalisms, and artifacts indicative of social standing, ethnic background and level of education often hold back persons otherwise eminently qualified to advance themselves in life. For this reason, speech, as a subject, has long been a part of the curricula in many schools, although in recent years education in this area has become, more and more, relegated to courses of study highly dependent on good speaking ability, such as radio, television, motion pictures and the theater.

Part of the problem here has been the difficulty of finding good voice instructors and the relatively high cost of the individualized instruction needed for a high degree of effectiveness in this area. In accordance with the invention, a specialized but highly effective speech training regimen is provided for application in the context of speech recognition software for receiving human language inputs in audio form to a microphone, analyzing the same in a

personal computer and outputting alphanumeric documents and navigation commands for control of the person computer.

In accordance with a present invention speech recognition is performed on a first computing device using a microphone to receive audible sounds input by a user into a first computing device having a program with a database consisting of (i) digital representations of known audible sounds and associated alphanumeric representations of the known audible sounds and (ii) digital representations of known audible sounds corresponding to mispronunciations resulting from known classes of mispronounced words and phrases. The method is performed by receiving the audible sounds in the form of the electrical output of a microphone. A particular audible sound to be recognized is converting into a digital representation of the audible sound.

The digital representation of the particular audible sound is then compared to the digital representations of the known audible sounds to determine which of those known audible sounds is most likely to be the particular audible sound being compared to the sounds in the database. A speech recognition output consisting of the alphanumeric representation associated with the audible sound

most likely to be the particular audible sound is then produced. An error indication is then received from the user indicating that there is an error in recognition. The user also indicates the proper alphanumeric representation of the particular audible sound. This allows the system to determine whether the error is a result of a known type or instance of mispronunciation. In response to a determination of error corresponding to a known type or instance of mispronunciation, the system presents an interactive training program from the computer to the user to enable the user to correct such mispronunciation.

The presented interactive training program comprises playback of the properly pronounced sound from a database of recorded sounds corresponding to proper pronunciations of the mispronunciations resulting from the known classes of mispronounced words and phrases.

In accordance with a preferred embodiment of the invention, the user is given the option of receiving speech training or training the program to recognize the user's speech pattern, although this is the choice of the user of the program.

In accordance with the invention, the determination of whether the error is a

result of a known type or instance of mispronunciation is performed by comparing the mispronunciation to the digital representations of known audible sounds corresponding to mispronunciations resulting from known classes of mispronounced words and phrases using a speech recognition engine.

It is anticipated that the inventive method will be implemented by having the database consisting of (i) digital representations of known audible sounds and associated alphanumeric representations of the known audible sounds and (ii) digital representations of known audible sounds corresponding to mispronunciations resulting from known classes of mispronounced words and phrases, generated by the steps of speaking and digitizing the known audible sounds and the known audible sounds corresponding to mispronunciations resulting from known classes of mispronounced words and phrases. The database will then be introduced into the computing device of many users after the generation by speaking and digitizing has been done on another computing device and transferred together with voice recognition and error correcting subroutines to the first computing device using CD-ROM or other appropriate data carrying medium.

It is also contemplated that mispronunciations are input into the database by actual speakers that have such errors as a natural part of this speech pattern.

In accordance with the invention, normalization to word, phrase and other sound models may be achieved by normalizing words or phrases to one of a plurality of sound durations. This procedure is followed with respect to all the word that phrase models in the database. When a word is received by the system, it measures the actual duration, and then normalizes the duration of the sound to one of the plurality of the preselected normalized sound durations. This reduces the number of items in the database against which the sound is compared and rated.

## BRIEF DESCRIPTION OF THE DRAWINGS

One way of carrying out the invention is described below in connection with the figures, it which:

Figure 1 is a block diagram illustrating a voice recognition program in accordance with the method of the present invention;

## DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

Referring to Figure 1, the system and method of the present invention may be understood. In accordance with the inventive method 10, a voice and error model is generated using subroutine 12. Subroutine 12 comprises a number of steps which are performed at the site of the software developer, the results of which are sent for example, in the form of a CD-ROM, other media or via the Internet, together with the software for executing voice recognition, to a user, as will be apparent from the description below. In accordance with the present invention, the inventive speech recognition method may be practiced on personal computers, as well as on more advanced systems, and even relatively stripped down lightweight systems, the referred to as subnotebooks and even smaller systems provided at the same have sound boards for interfacing with and receiving the output of a microphone. It is noted that quality sound board electronics is important to good recognition.

## SOFTWARE DEVELOPMENT PHASE

At step 14 a database of word models is generated by having speakers speak the relevant words and phrases into a microphone connected to the sound board of the personal computer being used to generate the database. In accordance with

the preferred embodiment of the invention, speakers who have been trained in proper speech habits are used to input words, phrases and sounds into the database at step 14. As the information is generated by the speakers speaking into microphones, attached to the sound boards in the computer, the information is digitized, analyzed and stored on the hard drive 16 of the computer.

In accordance with the present invention, relatively common pronunciation errors are also input into the system at step 18. In this specification the term "phoneme" is used to mean the smallest sound, perhaps meaningless in itself, capable of indicating a difference in meaning between two words. The word "dog" differs from "cog" by virtue of a change of the phoneme "do" pronounced "daw" and "co" pronounced "cah."

Thus, at step 18, the model generating speaker can speak a database of common phoneme errors into the microphone attached to the sound board of the computer to result in input of an error database into hard drive 16 of the computer. However, it is preferred that the phoneme errors are spoken by persons who in various ways make the pronunciation error as part of their normal speech paterns.

At step 20, the system is enhanced by the introduction into the database

contained on hard drive 16 of a plurality of exercise word models, selected for the purpose of training the speech of a user of the system. The same are input into the system through the use of a microphone and sound board, in the same way that the database of the language model was input into the system. Generally, a collection of word and/or phrase models is associated with each type of phoneme error. This is because if a person makes a speech pronunciation error of a particular type, it is likely that the same speaker makes certain other errors which have common characteristics with other pronunciation errors in the group. For example, a person who mispronounces the word... to sound like... is also likely to mispronounce the words...

Exercise phrase models are input at step 22. These exercise phrase models are stored by the system in hard drive 16. The exercise word models and the exercise phrase models input into the system at steps 20 and 22, respectively are associated in groups having common mispronunciation characteristics. The same are input into the system through the use of a microphone and sound board, in the same way that the database of the language model was input into the system.

In addition, in accordance with the present invention, it is recognized that computer errors may result in misrecognition of a particular error, mistaken acceptance of a mispronunciation, or mistaken rejection of a proper pronunciation. Accordingly, during the database generation session during which properly pronounced exercise word models or exercise phrase models or input into the system at steps 20 and 22, audio recordings of the same are also

stored on hard disk 16, to allow for playback of these proper pronunciations during use the program by a person performing speech recognition using the program. This provides for an audible cue to the user and allows the user to monitor the reliability of the system during the voice recognition and speech training process of the present invention.

In accordance with the invention, it is anticipated that there may be more than one mispronunciation associated with a particular word or phrase. Accordingly, at step 24, a plurality of typical mispronunciations are input into the system to create a database of exercise word error models in hard drive 16. The same are input into the system through the use of a microphone and sound board, in the same way that the database of the language model was input into the system.

Finally, the database of relatively common mispronunciation errors is completed at step 26 where the speaker generating that database speaks into the system to generate a plurality of exercise phrase error models. These error models are also input into the system through the use of a microphone and stored on hard drive 16.

In accordance with a preferred embodiment of the invention, the input of audible sounds into the system to generate the word error models at step 24 and the exercise phrase error models at step 26 is done using a speaker or speakers who have the actual speech error as part of their normal speech patterns. The same is believed to achieve substantially enhanced recognition of speech errors, although the same is not believed to be necessary to a functioning system.

In accordance with the preferred embodiment of the invention, the models stored on hard disk 16, and generated as described above may be recorded on a CD-ROM or other program carrying media, together with a voice recognition engine, such as that marketed by any one of a number of manufacturers such as IBM, Dragon Systems, and others. In accordance with a present invention, such a prior art speech recognition program may be used for both the purpose of recognizing words, recognizing mispronunciations and phoneme errors, together with the above described audio recordings of proper pronunciations, both during speech recognition operation training sessions.

In accordance with the invention, such software comprising the speech recognition engine, editing and training utilities, and database of word models, phrase models, vocal recordings, and error models may be supplied to the user for a one time fee and transported over a publicly accessible digital network, such as the Internet. Alternatively, the software may be made available for limited use for any period of time with charges associated with each such use, in which case the software would never be permanently resident on the computer of a user.

## USER TRAINING PROGRAM

When a user desires to use the inventive program, the software containing the program and the database is loaded into a personal computer and words are spoken into a microphone coupled to the sound board of the computer, in order to input the speech into the computer in the manner of a conventional speech

recognition program.

More particularly, as discussed above, after the system has proceeded through the performance of steps 14, 18, 20, 22, 24 and 26, and the speech recognition engine, editing and training utilities added, the system proceeds at step 28 to receive, through a microphone, speech to be recognized from a user of the program who has loaded the speech recognition engine, editing and training utilities, and database of word models, phrase models, vocal recordings, and error models onto the user's personal computer. In this respect, the operation of the speech recognition program of the present invention is substantially identical to other speech recognition programs presently on the market. More particularly, at step 30, a conventional speech recognition algorithm is applied to recognize audible sounds as the words which they are meant to represent.

The computer then outputs the recognized speech on the screen of the computer monitor, and the next phrase uttered by the user proceeds at step 30 through the speech recognition algorithm resulting in that speech also being displayed on the monitor screen. When the user notices that an error has occurred, he may use any one of a number of different techniques to bring up an error correction window at step 32. For example, he may simply double-click on the error, or highlight the erroneous recognition and hit a key dedicated to presentation of the error correction window.

User correction occurs at step 34. In typical programs, call up of the error correction window results in the presentation of a screen showing the

highlighted word, and suggesting, through the use of a menu, a number of alternatives which may be selected for double-clicking, in order to correct the error. If the problem word is not in the menu of alternatives, the user may type in the problem word or spell it out. After the system has been given the correct word by any of these means, the same is input into the system.

At this point, the call up of the error correction window at step 34 has indicated to the system that there is an error. While some errors are unrelated to pronunciation errors, many are. Once the user indicates the error, the system then proceeds at step 36 to determine whether the error made by the user is recognized as one of the speech errors recognized by the system. If it is, this information is determined at step 36. That the nature of the pronunciation error is then input into the system and logged at step 38. In this matter, the system keeps track of the number of errors of a particular type for the user by storing them and tallying them in hard drive 16.

In accordance with the present invention, it is contemplated that the speech training will not be triggered by a single mispronunciation. Instead, it is contemplated that the repeated instances of a single type of mispronunciation error will be tallied, and when a threshold of pronunciation errors is reached in the tally, only then will speech training be proposed by the appearance of the screen of a prompt window suggesting speech training. The same could take the form of a window having the words "The system has determined that it is likely that we can improve your recognition by coaching you. Would you like to speak to the speech coach?" The screen may also have a headline above the

question, such as "The coach wants to talk to you!" The screen will also have a button bar "OK" and another marked "Cancel", to give the user the opportunity to click on the "OK" button to start a training session, or to click on the "Cancel" button to cancel the speech coaching session.

It is also noted that other combinations of events may be used to trigger training. For example, if the particular mispronunciation detected is a very well-defined one, such as the almost uniform tendency of some speakers to mispronounce the word "oil" as "earl", the definiteness with which this error has been determined makes training relatively likely to be necessary, and the threshold can be lowered to, for example, one instance of that error being detected. In other cases, or in the general case, one may wish to set the threshold at three, five or even ten instances of the error before the "The coach wants to talk to you!" screen is presented to the user of the system.

Once a mispronunciation has been detected by the system as a result of the information input by the user to the user correction screen at step 34, the error correction algorithm operates in a manner identical to the speech recognition algorithm at step 30, except that the error correction algorithm checks the database of common phoneme errors input into the system by the software developer at step 18 and the exercise word error models and exercise phrase error models input at steps 24 and 26. In connection with this, it is noted that the so-called phoneme errors relate to particular sounds consisting of one syllable or less, while the phrase and word models are somewhat more general, as described herein.

Thus, if, at step 40 the system determines that the threshold number of errors in that class has not been reached, it sends the system back to step 28, where speech recognition proceeds. If, on the other hand, a predetermined number of errors of the same class have been detected by the system and logged at step 38, at step 40 the system is sent to step 42 where the above described "The coach wants to talk to you!" screen is presented to the user, who is thus given the opportunity to train his voice.

If the user declines the opportunity to train at step 42, he is given the opportunity to train the database at step 43. If he declines that opportunity also, the system is returned to step 28, where, again, speech recognition proceeds.

However, if he accepts the opportunity to train the database, the system proceeds to step 45, where the database is trained in the same manner as a conventional speech recognition processing program.

In the other case, at step 42, when the user decides to accept speech training, the system proceeds to step 44, where the determination is made as to whether the particular error is an error in the pronunciation of a word or what is referred to herein as a phrase. By "phrase" in this context, is meant at least parts from two different words. This may mean two or more words, or the combination of one or more words and at least a syllable from another word, and most often the end of one word combined with the beginning of another word, following the tendency of natural speakers to couples sounds to each other, sometimes varying their stand-alone pronunciation. If, at step 44 the system determines

that the mispronunciation is the mispronunciation of a word, the system is sent to step 46 where the system retrieves from memory words which have the same or similar mispronunciation errors.

As noted above, these words have been stored in the system, not only in the form of alphanumeric presentations, but also in high-quality audio format. The object of the storage of the high-quality audio sound is to provide for audible playback of the words in the training dialog screen.

The words retrieved at step 46 are also presented on-screen in alphanumeric form to the user and the user is invited to pronounce the word. If the word is pronounced properly, this is determined at step 48. If there is no error, the system proceeds to step 50 where the system determines whether there are two incidences of no error having occurred consecutively. If no error has occurred twice consecutively, the system is returned to act as a voice recognition system at step 28. If no error has occurred only once, at step 50 the system is returned to the training dialog screen at step 46 and the user is invited to pronounce the same or another word having the same type of mispronunciation to ensure that the user is facet word correctly. Once the user has pronounced words twice in a row without errors, the user is returned at step 50 to the voice recognition function.

However, where an error has been detected at step 48, the system proceeds to step 50 to where an instruction screen telling the user how to make the sound, with physical instructions on how to move the muscles of the mouth and tongue

to achieve the sound is presented to the user.

The screen allows for the incorporation of more creative speech training approaches such as the Lessac method described in *The Use and Training of the Human Voice -A Bio-Dynamic Approach to Vocal Life,* Arthur Lessac, Mayfield Publishing Co. (1997). In this technique the user is encouraged to use his "inner harmonic sensing." This enhances the description of a particular sound by having the user explore how the sound affects the user's feelings or encourages the user to some action.

In an illustrative example, the Lessac method teaches the sound of the letter "N" by not only describing the physical requirements but also instructs the user to liken the sound to the "N" in violin and to "Play this consonant instrument tunefully." This screen also has a button which may be clicked to cause the system to play back the high-quality audio sound from memory, which was previously recorded during software development, as described above.

The system may also incorporate interactive techniques. This approach presents the user with a wire frame drawings of a human face depicting, amongst other information, placement of the tongue, movement of the lips, etc. The user may interactively move the wire frame drawing to get a view from various angles or cause the sounds to be made slowly so that the "facial" movements can be carefully observed.

The user is then invited to say the sound again, and at step 54, the user says the

word into the microphone which is coupled to the computer, which compares the word to the database for proper pronunciation at determines whether there is an error in the pronunciation of the word at step 56.

If there is error, the system is sent back to step 46 where, again, the word is displayed and the user invited to say the word into the machine to determine whether there is error, with the system testing the output to determine whether it should proceed to speech recognition at step 28, when the standard of two consecutive correct pronunciations has been reached. If there is no error at step 56, however, the tally is cleared and the system proceeds to step 28, where normal speech recognition continues.

If, at step 44 the system determines that the mispronunciation is the mispronunciation of a phrase, the system is sent to step 58 where the system retrieves from memory phrases which have the same or similar mispronunciation errors.

As noted above, these phrases have been stored in the system, not only in the form of alphanumeric presentations, but also in high-quality audio format. The object of the storage of the high-quality audio sound is to provide for audible playback of the words in the training dialog screen.

The words retrieved at step 58 are also presented on-screen and alphanumeric form to the user at the user is divided to pronounce the word. If the word is pronounced properly, this is determined at step 60. If there is no error, the

system proceeds to step 62 where the system determines whether there are two incidents of no error having occurred. If no error has occurred twice, the system is returned to act as a voice recognition system at step 28. If no error has occurred only once, at step 62 the system is returned to the training dialog screen to at step 58 and the user is invited to pronounce the same or to the word having the same type of mispronunciation to ensure that the user is facet word correctly. Once the user has pronounced words twice in a row without errors, the user is returned at step 62 to the voice recognition function.

However, where an error has been detected at step 60, the system proceeds to step 62 to where an instruction screen telling the user how to make the sound, with physical instructions on how to move the muscles of the mouth and tongue to achieve the sound is presented to the user as well as any other techniques such as the Lessac method described herein above.

This screen also has a button which may be clicked to cause the system to playback the high-quality audio sound from memory, which was previously recorded during software development, as described above.

The user is then invited to say the sound again, and at step 66, the user says the phrase into the microphone which is coupled to the computer, which compares the word to the database for proper pronunciation and determines whether there is an error in the pronunciation of the word at step 68.

If there is error, the system is sent back to step 58 where, again, the word is displayed and the user invited to say the word into the machine to determine whether there is error, with the system testing the output to determine whether it should proceed to speech recognition at step 28, when the standard of two consecutive correct pronunciations has been reached. If there is no error at step 68, however, the tally is cleared and the system proceeds to step 28, where normal speech recognition continues, the training session having been completed.

While an illustrative embodiment of the invention has been described together with several alternatives for various parts of the system, it is, of course, understood that various modifications will be obvious to those of ordinary skill in the art. Such modifications are within the spirit and scope of the invention, which is limited and defined only by the following claims.